THE AVERAGE READING VOCABULARY; AN APPLICATION OF BAYES'S THEOREM.

By WARREN WEAVER, University of Wisconsin.

The rule for the computation of *a posteriori* probabilities was first developed by an English clergyman, T. Bayes, and was published after his death in the *Philosophical Transactions* for 1763. The careless application of this rule has led to many paradoxical results,[1] in consequence of which some mathematicians would abandon the rule entirely. Among this number may be mentioned Mr. J. Bing, a Danish actuary, the late Dr. T. Thiele, and especially Professor Chrystal, whose advice is to "bury the laws of inverse probability decently out of sight." The problem herein stated and solved may be of interest since it clearly emphasizes the point the neglect of which has led to incorrect results, since it shows what great allowances may sometimes be made in the *a priori* probabilities of existence and still allow us to change our view in regard to a statistical result, and since it is an answer to a specific case of that interesting question, to what degree does new experimental evidence justify us in modifying previously held opinions.

1. **Statement of the theorem.** Some one of the mutually exclusive causes $A_1, A_2, \cdots A_n$ is to produce an event. When the result is not known (*i.e.*, before the event occurs) the existence probability for each cause is $\pi_1, \pi_2, \cdots \pi_n$ (*i.e.*, the *a priori* probability of the existence of each cause). The event in question occurs. The cause A_1, when it is known to act, gives the productive probability p_1, etc. Then the *a posteriori* probability that the cause A_r produced the event is

$$P_r = \frac{p_r\pi_r}{p_1\pi_1 + p_2\pi_2 + \cdots + p_r\pi_r + \cdots + p_n\pi_n}. \tag{1}$$

If the event in question is able to occur in two alternative ways one of which we call "successful," and the other of which "unsuccessful"; and if, further, the productive probability that the cause A_r produce the event successfully be ω_r, then if the event occurs successfully m times in k trials the *a posteriori* probability that the cause A_r has been the one to act is

$$P_r = \frac{\pi_r\omega_r^m(1-\omega_r)^{k-m}}{\Sigma\pi_i\omega_i^m(1-\omega_i)^{k-m}} \quad (i = 1, 2, \cdots n). \tag{2}$$

The theorem may assume a third form in problems of such a nature that the different causes A_r may be considered different stages of a continuously changing complex. In this case the quantities involved in the formula become definite integrals.[2]

[1] For example, Bing's Paradox: "If among a large group of S equally old persons we have observed no deaths during a full calendar year, then another person of the same age outside the group is certain to die inside the calendar year" quoted from *The Mathematical Theory of Probabilities*, by A. Fisher. Volume 1, New York, 1915, p. 75.

[2] A. Fisher, *l.c.*, p. 67.

To those not entirely familiar with the theorem an example may make the statement of it more clear. Suppose that we have an urn filled with black and white balls in unknown proportion, and that our *a priori* estimate of the existence probability that there are x white and $(b - x)$ black is π_x (b being the total number of balls in the urn). Suppose that we draw k balls from the urn, returning each, and find that of these m are white and $k - m$ black. What is then the most probable mixture in the urn? It will not be the one originally most probable, that is, the one for which π_x is a maximum; nor will it be the one suggested by the drawing,[2] but will obviously be some mixture intermediate between these two. It will be, in fact, a mixture of x^* white and $(b - x^*)$ black, where x^* is that value of x for which

$$P_x = \frac{\pi_x \left(\frac{x}{b}\right)^m \left(\frac{b-x}{b}\right)^{k-m}}{\Sigma \pi_{x'} \left(\frac{x'}{b}\right)^m \left(\frac{b-x'}{b}\right)^{k-m}} \qquad (x' = 1, 2, \cdots n) \tag{3}$$

is a maximum. In case π_x is given, by experiment, judgment, or calculation, for a certain finite number of values of x, we might determine this value x^* by plotting to any scale whatsoever, and noting the value of x corresponding to the highest point on the curve. If we should later wish the vertical scale of this curve we could most easily determine it from the fact that the area under it must equal unity.

2. **Statement of the problem.** A test has been devised by the department of educational psychology at the University of Wisconsin to determine a person's reading vocabulary. The process consists of taking at random 200 words from the dictionary, and having a person decide with how many of the 200 words he is familiar—say 117. Then the value of this person's reading vocabulary is taken as

$$(117/200)\ 104{,}000$$

or about 61,000. (104,000 being the approximate number of words in the dictionary.)

The scheme has been found to give, as the result of about five hundred tests by university students, the value given above. This value is far in excess of previous estimates, the general opinion before this test being, according to Professor Starch, that the correct figure was in the neighborhood of 25,000. We wish to investigate whether this process gives us a sound basis for raising our previous estimate of 25,000 to 61,000, and if not, what our answer should be.

Since, as will appear later, the *a priori* probabilities of different estimates have an important effect upon the solution of the problem it is necessary to inquire as carefully as the nature of the existing information will permit into the basis and reliability of this estimate of 25,000 words. Unfortunately the information is vague, but it appears to be an average opinion of those who had been interested in the matter, rather than a definite statistical result. There may have been, to be sure, some numerical method, however unsatisfactory, by which the

[2] Unless, of course, these two mixtures happen to be the same.

estimates were arrived at; or it may be that an investigator with a great deal of experience along this line would venture an estimate based upon intuition alone. It is evident as a mere matter of common sense that if this estimate of 25,000 words were formed upon the basis of one application of this test itself, and the next application of the test furnished an estimate of 61,000, one would not be justified in completely discarding the old estimate for the new. It is obviously a matter of the comparative reliability of the new and old judgments, which comparison is made accurately by the theorem stated.

3. **Solution of the problem.** The actual method used in taking the sample was to take the first word on the kth page of a Webster's *Unabridged Dictionary*, k having such a value that the method would result in a sample of 200. There being so many words to pick from, it is evident that it is immaterial whether or not we consider that we return each word after its drawing: a consideration which in other cases might be important. It seems likely, on an intuitive basis, that if the same person performed the test several times with different samples, or if it were performed with several persons and the same sample, results would be obtained that would vary widely, especially since 200 seems a small sample from a group of 104,000. It should be emphasized therefore that the datum which we use is the average of over five hundred results from different persons. And some knowledge of the variability of these results is important in making an estimate of the stability of the average. The following frequency table gives us an estimate of the variability in the results obtained from a typical group of fifty students, using the same sample of words. It is on the basis of 100 words rather than 200 since, as a matter of proceedure in making the test, the whole list was split up into two lists of 100 each, and the score kept for each separately.

No. of Words Known.	No. of Students.	No. of Words Known.	No. of Students.
46	0	61	1
47	0	62	3
48	1	63	2
49	1	64	3
50	2	65	0
51	0	66	1
52	3	67	3
53	1	68	4
54	2	69	0
55	3	70	5
56	1	71	0
57	3	72	1
58	4	73	0
59	1	74	0
60	5	⋮	—
			50

The result shown is typical of all obtained, a mean variation of approximately five being found among all the students tested. We conclude therefore that 117 is an average of considerable stability, coming as it does from a group of over 500 tests which show a relatively small variability. Unfortunately the data are not available as to how many sets were used: the errors of sampling in the dictionary would be reduced in the approximate ratio of 1 to $\sqrt{n}$ where n is the number of sets used. However the accuracy of any one list as expressed by the mean variation between two lists of 100 words each is between two and three words. We are led then to this conclusion: that if we imagine the hypothetical case of "the average" university student examining a perfectly fair sample of 200 words it seems reasonable to assume that he would find among these 117 that he would know. And any conclusion we may draw from this hypothetical case will have a reliability of approximately $\sqrt{500} = 22.4$ times as much as it would have if it actually came from but a single trial. The fact that 200 is indeed a small sample will later appear in this, that the most probable mixture, even though it may be many times more probable than other mixtures not in the immediate neighborhood of the most probable one, is, as a matter of fact, a mixture whose probability is very small. This apparent paradox is often met with in problems involving large numbers.

A mere change in the wording of the problem makes the application of equation (2) evident. We have an urn filled with b (= 104,000) words in unknown proportion of "white" (known) words to "black" (unknown) words. From this urn we draw 200 and find 117 white and 83 black. In other words the event in question occurs 200 times in one of two alternative ways, it occurring 117 times in the way which we may call successful. The probability that the dictionary contains a mixture of x known and $b - x$ unknown words is then

$$P_x = \frac{\pi_x \left[\frac{x}{b}\right]^{117} \left[\frac{b-x}{b}\right]^{83}}{\sum_1^n \pi_x \left[\frac{x}{b}\right]^{117} \left[\frac{b-x}{b}\right]^{83}} \qquad (4)$$

or

$$P_x = K\pi_x \left[\frac{x}{b}\right]^{117} \left[\frac{b-x}{b}\right]^{83}, K \text{ being a constant.} \qquad (5)$$

We see at once that it is impossible to determine the value of x for which this expression is a maximum without a knowledge of the character of the term π_x. This is exactly the point at which errors often creep into applications of the theorem. It is often assumed from the logical principle of insufficient reason that π_x is a constant: in other words since we know too little to form a judgment it is assumed that the *a priori* probabilities of all causes are equal. The principle of insufficient reason leads, however, to notoriously paradoxical results.

For the problem here considered, however, while the theoretically correct result cannot be obtained without a knowledge of π_x, we can easily show that for practical purposes our knowledge concerning it may be very limited and still sufficient.

Consider the curve

$$P'_x = \left[\frac{x}{b}\right]^{117}\left[\frac{b-x}{b}\right]^{83} \tag{6}$$

where $b = 104{,}000$. This curve has its maximum value when

$$x = (117/200)\,b = 61{,}000 \tag{7}$$

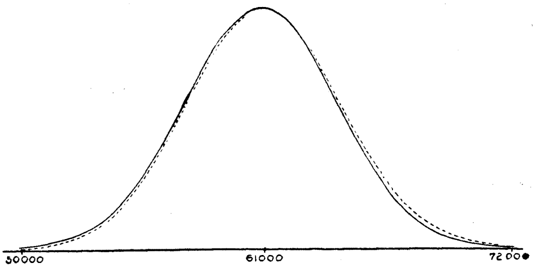
and the maximum value is equal to

$$P'_{61000} = 1.137 \times 10^{-59}. \tag{8}$$

However for $x = 25{,}000$

$$P'_{25000} = 2.207 \times 10^{-83}.$$

To obtain the value of x for which P_x of equation (5) is a maximum we have to multiply every ordinate of the curve P_x' by π_x (the existence probability of a mixture of x known and $b - x$ unknown words), and then take the value of x corresponding to the highest point on this new curve. Therefore unless the *a priori* existence probability of a mixture containing 25,000 known words exceeds the *a priori* existence probability of a mixture containing 61,000 known words in the ratio of 0.5×10^{24} to 1 it is evident that the *a posteriori* probability of a mixture characterized by $x = 61{,}000$ is greater than the *a posteriori* probability of a mixture in which x is only 25,000. In fact we have



$$\frac{P_{61000}}{P_{25000}} = \frac{K\pi_{61000} \times 1.137 \times 10^{-59}}{K\pi_{25000} \times 2.207 \times 10^{-83}} = 0.5 \times 10^{24}\frac{\pi_{61000}}{\pi_{25000}}, \tag{10}$$

which is greater than unity as long as

$$\pi_{25000} < 0.5 \times 10^{24} \times \pi_{61000}. \tag{11}$$

While this may convince us that the previous estimate of 25,000 is to be discarded it may not convince us that an estimate of, say, 50,000 is not as good a new estimate as that of 61,000 which the test indicates. Let us therefore consider the numerical magnitude of the ratio of the *a posteriori* probability of a mixture for which $x = 61{,}000$ to the *a posteriori* probability of a mixture for which $x = 50{,}000$. We have

$$\frac{P_{61000}}{P_{50000}} = \frac{\pi_{61000}}{\pi_{50000}}\left[\frac{61}{50}\right]^{117}\left[\frac{43}{54}\right]^{83} = 77\,\frac{\pi_{61000}}{\pi_{50000}},$$

which is greater than unity unless the *a priori* probability of a mixture of 50,000 known words exceeded the *a priori* probability of a mixture of 61,000 known words by the factor 77. Since neither of these figures, 50,000 and 61,000, is in any way special before the test is made there would seem no justifiable basis for considering one more probable than the other in any such ratio as that just found. The answer to our problem would then be that 61,000 is the most probable answer for the number of words in the dictionary, this conclusion being reached regardless of the character of π_x outside of the one restriction stated in (11). It is understood, of course, that the character of the term π_x in the neighborhood of $x = 61{,}000$ might shift the most probable value slightly, but π_x would surely be changing very slowly for values of x in this vicinity, and the shift would be therefore very small, and negligible for practical purposes. Although the question of whether equation (11) states a reasonable restriction upon the character of π_x is primarily one for educators to settle it would certainly seem sensible to assume that it does. We should surely agree that the previous results were not sufficiently well established that we could consider them, *a priori,* 0.5×10^{24} times as likely to be true as any other result. We must say "any other" result since our estimate of the *a priori* probabilities, it being independent of the result of the test and therefore for psychological reasons best formed before the test takes place, could attach no special importance to the figure 61,000—a number which is not known until after the test is performed. All we could say might be, for example, that we consider a result lying between, say, 20,000 and 30,000 one hundred times more likely than a result lying outside this band, and that we consider it certain that the actual mixture contains more than 5,000 and less than 90,000 words that the average student knows. Such an assumption, coupled with the fact that the total area under the curve $y = \pi_x$ must be unity gives

$$\left.\begin{aligned} \pi_x &= 0 & 0 &\leq x < 5000 \\ &= 9.302 \times 10^{-7} & 5000 &\leq x < 20000 \\ &= 9.302 \times 10^{-6} & 20000 &\leq x < 30000 \\ &= 9.302 \times 10^{-7} & 30000 &\leq x < 90000 \\ &= 0 & 90000 &\leq x \leq 104000 \end{aligned}\right\} \qquad (12)$$

Then we have P_x given by the full line on the graph.

The vertical scale is again obtained from the fact that the area must equal unity. The probability of the most probable mixture is found to be 9.946×10^{-5}, a very small probability as was earlier suggested would be the case. The values of P_x outside the range shown are too small to be indicated.

It is to be especially noted that this curve will approximately represent P_x *whatever the assumption concerning* π_x, only provided, say, that

$$\pi_{25000} < .5 \times 10^{20}\, \pi_{61000} \qquad (13)$$

This condition is slightly more stringent than (11), and insures that P_{25000} shall be less than one ten thousandth of P_{61000}, and therefore negligible. It is easily shown that in a small neighborhood of $x = (117/200)\,b$ the equation

$$P_x = K\pi_x \left[\frac{x}{b}\right]^{117} \left[\frac{b - x}{b}\right]^{83} \tag{5}$$

is equivalent to

$$P_x = K\pi_x P'_{61000}\, e^{[-(117+83)3/2][(1/83)(1/117)]\epsilon^2}, \tag{14}$$

where

$$\epsilon = (117/200)b - x. \tag{15}$$

This equation reduces to

$$P_x = Ce^{-411.9\epsilon^2} \tag{16}$$

if we assume, as we have done, that π_x is constant in a small range about $x = 61{,}000$. The constant C is evaluated by means of the fact that when $\epsilon = 0$, P_x must be the probability of a mixture containing 61,000 known words, which probability we have previously calculated. Then

$$C = 9.946 \times 10^5. \tag{17}$$

The size of the coefficient of ϵ^2 in equation (16) indicates clearly how rapidly probabilities diminish in the neighborhood of the most probable result. The approximation of equation (16) to equation (5) is shown on the graph. The dotted line is the graph of equation (16).

We know from Bernoulli's theorem that if the dictionary actually consisted of x^* words which we could characterize as known, and $b - x^*$ which would accordingly be unknown, as we take more and more samples from it the estimate formed from these samples must approach the value x^*. We have, indeed, for the ratio of the *a posteriori* probability of a mixture containing x known words to the *a posteriori* probability of a mixture containing 61,000 known words

$$\frac{P_x}{P_{61000}} = \frac{\pi_x}{\pi_{61000}} \left[\frac{x}{61000}\right]^m \left[\frac{104000 - x}{43000}\right]^{k-m},$$

where m known words have appeared in a total sample of k. If $m/k - m$ equals 61,000/104,000 the above ratio has its maximum when $x = 61{,}000$, in which case it is obviously unity. For any other value of x this ratio may be greater than unity for a given k, but must become and remain smaller than unity, as k increases indefinitely whatever the (finite) ratio of π_x to π_{61000}. The truth of this statement is obvious from the form of the above equation. It thus appears that the *a priori* probability is vanishingly unimportant as the number of trials increases. If it were the case that 500 tests had been performed with the invariable result of 117 known out of 200 chosen in each test we would have $k = 100{,}000$ and $m = 58{,}500$. It is then clear that the above ratio would be exceedingly small for any value of x other than 61,000 practically independently of the ratio π_x/π_{61000}.

It is, however, not strictly admissible to make such a calculation in our case. For one thing the result in the 500 tests differed, even though with surprisingly small variability. Even more important than this, however, is the fact that we cannot in all strictness compare our problem with the analogous urn problem in case the test is used on more than one person, as it of course was. For the content of the dictionary, from our point of view, actually changes with the observer, depending, as it does, upon how many words *he* knows. This is an added source of variability, over and beyond that which would occur due to the ordinary errors of sampling. The result of the above paragraph is still qualitatively applicable.

Our final conclusion is, then, that the experimental evidence of the test completely justifies us in abandoning the old result and accepting the new: and that, moreover, probabilities of mixtures in the immediate neighborhood of the most probable mixture themselves follow the normal Gaussian law as given by equation (16).

A GRAPHICAL AID IN THE STUDY OF FUNCTIONS OF A COMPLEX VARIABLE.

By NORMAN MILLER, Queen's University.

The impossibility in three dimensions of representing graphically a function of a complex variable makes it necessary for the student to call on his imagination in other ways in order to realize the properties of these functions. Two methods are common in the geometrical theory of functions. One is to represent in two different planes or in two Riemann surfaces the variables z and w and to study the correspondence between the points of the two planes or surfaces, which is determined by the relation $w = f(z)$. The second method, which does much to illuminate the subject for the beginner, is to represent in one plane both the independent and dependent variables and to interpret the transformation kinematically as a flow of the points in the plane.[1]

A complete graph of the function $w = f(z)$ or $u + iv = f(x + iy)$ consists of a 2-dimensional manifold in space of four dimensions. Nevertheless the student, in his effort to visualize the function, thinks instinctively of a surface spread out over the plane of z. Such a surface is actually determined by taking for a third coördinate the absolute value of $f(z)$. Calling the third coördinate ζ the equation of the surface is

$$\zeta = \sqrt{[u(x, y)]^2 + [v(x, y)]^2},$$

only the positive square root being taken. In this representation all points on a circle of center 0 in the w-plane yield the same ordinate ζ. It is, in fact, by making no distinction among the points of such a circle that we are able to pass from a two-way spread in four dimensions to an actual surface in three dimensions.

It is interesting to enquire what properties of the function $f(z)$ are exhibited

[1] See in this connection an article by Cole, *Annals of Mathematics*, vol. 5, June, 1890.